Journal of Biometrics & Biostatistics

ISSN: 2155-6180

Dai et al., J Biom Biostat 2012, 3:7
DOI: 10.4172/2155-6180.1000e117

Editorial Open Access

Integrating P-values for Genetic and Genomic Data Analysis

Hongying Dai[1]*, Richard Charnigo[2], Tarak Srivastava[3], Zohreh Talebizadeh[4] and Shui Qing Ye[4,5]

[1]Research Development and Clinical Investigation, Children's Mercy Hospital, 2401 Gillham Road, Kansas City, MO, 64108, USA
[2]Department of Statistics and Biostatistics, University of Kentucky, 725 Rose Street, Lexington, KY, 40536, USA
[3]Section of Nephrology, Children's Mercy Hospital, 2401 Gillham Road, Kansas City, MO, 64108, USA
[4]Division of Genetics Research, Department of Pediatrics, Children's Mercy Hospital, 2401 Gillham Road, Kansas City, MO, 64108, USA
[5]Department of Biomedical and Health Informatics, University of Missouri-Kansas City School of Medicine, 2464 Charlotte Street, Kansas City, MO 64108, USA

Rapid developments in molecular technology have led to evolution in Biostatistics and Bioinformatics, to identify genetic variations associated with complex traits. A large amount of information becomes accessible to investigators through Genome Wide Association Studies (GWAS), gene expression arrays, whole genome sequencing and other technologies.

The increase of variants requires more statistical testing to be conducted in analyses, which poses a "curse of dimensionality" to multiple testing correction methods. For instance, false discovery rate (FDR) and its extended methods are commonly used to adjust multiple individual tests, in order to control the family wise Type I error [1,2]. Unfortunately, in large-scale hypothesis testing, these methods tend to yield low power to detect risk factors.

Global testing (also named omnibus testing) of p-values from numerous individual tests may combine evidence, and turn dimensionality from a curse into rich information. From a systems biology perspective, genes, cells, tissues and organs function as a system through metabolic networks and cell signal networks. In non-Mendelian inheritance such as complex disorders, a subset of variants may jointly confer moderate effects in mediating molecular activities. As a result, signals may not be significant in single marker-single trait analysis, but many such values from related genes might provide valuable information on gene function and regulation.

The global test is designed to evaluate the pattern (distribution) of p-values, instead of choosing p-values less than an arbitrary threshold. Therefore, this method has the potential to identify multiple genes with small effects. Assuming that all individual tests are independent and arise from genes with no effects, p-values are identically and independently distributed as *Uniform*(0,1). Taking this as a null hypothesis for the pattern of p-values in the global test, one can assess whether p-values, especially small p-values, are generated by chance. The global test of p-values is robust and can be applied to p-values from a t-test, an ANOVA, a linear mixed model, and so forth. Multiple simulation studies and case studies have demonstrated that the approach usually has sufficient power to detect signals of genetic association from a group of genes.

Methods

Combination of p-values into a sum or product has long been used by evolutionary biologists in meta-analysis [3]. Many methods can be expressed in the form of $T=\sum_i H(p_i)$, where p-values might first be transformed by a function *H*. Early researchers had been exploring a raw sum of p-values and sums with various transformations, including log transformation, inverse normal transformation, inverse gamma transformation, logit transformation, and count of p-values less than a threshold, etc. Some classic methods include Fisher's method [4], Z-test [5], and Lancaster's procedures [6]. Extensive Monte Carlo comparisons have been conducted for independent [7], and correlated [8] p-values. The classic methods yield simple limiting distributions when p-values follow the identical and independent uniform distribution, under the global null hypothesis. One can also combine p-values using the product method $T=\prod_i H(p_i)$ [9,10]. By taking log-transformation on the product of p-values, the product method becomes a special case of sum of log-transformed p-values $ln(T)=\sum_i ln(H(p_i))$.

Order-based approaches are another category of global testing for p-values. Tippett's procedure is to assess the minimal p-value. Simulation studies show that this approach has well controlled Type I error for both independent and correlated data, but will reduce power to identify multiple genes with small effects [8]. Wilkinson extended Tippett's procedure to the *k* smallest p-values. By expanding $(\alpha+(1-\alpha))^m$, where *m* is the total number of individual tests, tables of the incomplete beta function can be used to obtain the probability of tests with p-values less than α [11]. Furthermore, empirical distributions of p-values can be calculated and compared to the uniform distribution. These tests include the positive-side Kolmogorov-Smirnov test, the positive-side Cramer-von Mises test, the newly developed order-based approach that accounts for ordering of p-values under the alternative hypothesis [12], and the higher criticism method to detect sparse signals [13].

Current Trends

Recent developments have focused on introducing weight functions and truncation to increase power, as well as on developing global tests for genetic analysis. For instance, a rank truncated method that combines the first *k* ordered p-values and a truncated product method, that combines p-values that are smaller than a specified threshold, have recently been developed and applied in large scale genomics experiments [14]. Later, an adaptive rank truncated product method was proposed and applied in GWAS [15]. By Yu et al. [15], permutation testing was used to determine the optimal number of *k* smallest p-values for a product test. In Hess and Iyer [16], Fisher's method was extended to Affymetrix gene expression arrays and shown to be a suitable diagnostic tool for exploratory analysis of microarray data. The combined p-value method was shown to be favorable versus competing methods through validated microarray data analyses.

***Corresponding author:** Hongying Dai, Research Development and Clinical Investigation, Children's Mercy Hospital, 2401 Gillham Road, Kansas City, MO, 64108, USA, E-mail: hdai@cmh. edu

Received November 02, 2012; **Accepted** November 05, 2012; **Published** November 12, 2012

Citation: Dai H, Charnigo R, Srivastava T, Talebizadeh Z, Ye SQ (2012) Integrating P-values for Genetic and Genomic Data Analysis. J Biom Biostat 3:e117. doi:10.4172/2155-6180.1000e117

Efforts have also been made to cope with complex correlations among p-values. In expression quantitative trait loci (eQTL) analysis to identify genotype and phenotype associations [17], researchers have observed strong correlations among multiple tests due to linkage disequilibrium and functional interactions among single nucleotide polymorphisms (SNPs). To address this issue, Fisher's method was modified to incorporate correlations among p-values, and then a Satterwhite's approximation was used to derive the limiting distribution of the test statistic, under the global null hypothesis. Similarly, the weighted Z-test has been modified to include correlations and has been applied in shared controls designs in GWAS [18].

Modeling p-values using analytic distributions also starts to show promise. A beta mixture model has been proposed to model p-values that might come from a combination of null and alterative hypotheses for individual genes. Then, a modified likelihood ratio test and a D-test are proposed to test homogeneity in the mixture model [19]. In Dudbridge and Koeleman [20], extreme-value distributions for fixed numbers of combined evidence and a beta distribution for the most significant evidence are shown to be accurate and efficient for large exploratory studies. Analytic modeling may provide a deeper level of insight into properties of p-values. For instance, a mixture model of p-values may not only suggest the existence of overall signals, but also measure the proportion of variants associated with a phenotype, as well as the strength of association effects.

Below we describe two major trends in application of the combined evidence approach to complex genetic data analysis.

Filtration of Variants with No Association

Global tests can filter out genes with no association and direct researchers to a smaller part of the genome [19]. Filtration is a critical process in current genetic data analyses to remove noises, irrelevant variants and weak signals. Removing genes using arbitrary cutoff values (such as fold change>1. 5 or p-value<0. 05) might increase bias in gene selection. We advocate incorporating global tests into a gene filtration process. Essentially, one can group genes into gene sets based on biological information, pathway or functional network etc. Global tests of p-values will then be performed in the various gene sets to detect whether overall signals exist. Gene sets with no overall signals will be removed, which will greatly reduce the dimensionality.

A global test of p-values can also be used to select the optimal number of genes for a final analysis. For instance, if an auxiliary measure can be used to rank the genetic variants and this auxiliary measure is independent of the global test, then the global test can be used to find a cutoff for the auxiliary measure and select the optimal number of genetic variants for the final analysis. In MDR analysis (the method for gene-gene interaction), several filtration algorithms (such as SURF [21], TuRF [22], and Relief F [23]) have been developed to rank SNPs based on efficiency and redundancy. Then, global tests can be used to determine the optimal cutoff points for these measures and select the optimal number of genes. The global test and ReliefF combined filtration approach has been applied to a candidate gene study of drug response in Juvenile Idiopathic Arthritis, and has identified gene-gene interaction in the folate pathway [8].

Two-stage Reversed Pathway Analysis

Pathway analysis is a field of study to detect a wide range of molecular entities which regulate specific cell functions, metabolic processes and biosynthesis. In Traditional Pathway Analysis (TPA), adjusted cutoffs of fold changes/p-values are being used to select significant individual genes (step 1). Next, it will be tested whether significant individual genes are over represented in pathways (step 2). However, the bias and random error in individual gene selections may severely impact subsequent steps of TPA. We suggest incorporating global testing into pathway analysis and reversing the aforementioned two steps by first detecting significant pathways, and then detecting significant genes in the significant pathways, as illustrated in figure 1. By switching to this omnibus testing based pathway analysis (OPA), the number of multiple tests is dramatically reduced from ~10^5 to ~10^2.

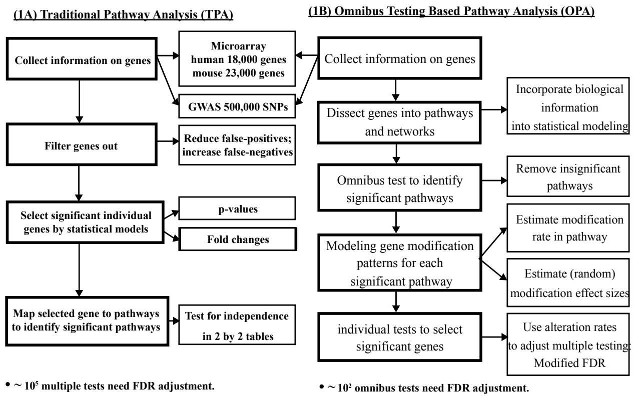
(1A) Traditional Pathway Analysis (TPA)
Collect information on genes
Microarray human 18,000 genes mouse 23,000 genes
GWAS 500,000 SNPs
Filter genes out
Reduce false-positives; increase false-negatives
Select significant individual genes by statistical models
p-values
Fold changes
Map selected gene to pathways to identify significant pathways
Test for independence in 2 by 2 tables
• ~ 10⁵ multiple tests need FDR adjustment.
(1B) Omnibus Testing Based Pathway Analysis (OPA)
Collect information on genes
Dissect genes into pathways and networks
Incorporate biological information into statistical modeling
Omnibus test to identify significant pathways
Remove insignificant pathways
Modeling gene modification patterns for each significant pathway
Estimate modification rate in pathway
Estimate (random) modification effect sizes
individual tests to select significant genes
Use alteration rates to adjust multiple testing Modified FDR
• ~ 10² omnibus tests need FDR adjustment.

Figure 1: Comparisons between traditional pathway analysis (TPA) and omnibus test based pathway analysis (OPA).

Conclusions

Fisher's method was shown to be asymptotically Bahadur optimal and efficient, assuming p-values are independent. However, there is no uniformly most powerful method of combining p-values. Moreover, accounting for correlations among p-values represents a major challenge to applying global methods that were originally designed based on independence assumptions. Using methods that are designed for correlated data will effectively prevent inflation of Type I error due to complex correlation structures. More ground-breaking theoretical works are needed to develop global tests of p-values that account for such correlation structures.

References

1. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Methodol 57: 289-300.

2. Cheng C, Pounds S (2007) False discovery rate paradigms for statistical analyses of microarray gene expression data. Bioinformation 1: 436-446.

3. Hedges LV, Olkin I (1985) Statistical methods for meta-analysis. (1st edn), Academic Press, San Diego.

4. Fisher R (1932) Statistical methods for research workers. Oliver and Boyd, Edinburgh.

5. Stouffer S, DeVinney L, Suchmen E (1949) The American Solder: Adjustment During Army Life, Vol 1, Princeton University Press, Princeton, NJ.

6. Lancaster HO (1961) The combination of probabilities: an application of orthonormal functions. Aust J Stat 3: 20-33.

7. Loughin TM (2004) A systematic comparison of methods for combining p-values from independent tests. Comput Stat Data Anal 47: 467-485.

8. Dai H, Bhandary M, Becker M, Leeder JS, Gaedigk R, et al. (2012) Global

tests of p-values for multifactor dimensionality reduction models in selection of optimal number of target genes. BioData Min 5: 3.

9. Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS (2002) Truncated product method for combining P-values. Genet Epidemiol 22: 170-185.

10. Dudbridge F, Koeleman BP (2003) Rank truncated product of P-values, with application to genomewide association scans. Genet Epidemiol 25: 360-366.

11. Bryan W (1951) A statistical consideration in psychological research. Psychol Bull 48: 156-158.

12. Ori D (2011) Combining p-values using order-based methods. Comput Stat Data Anal 55: 2433-2444.

13. Donoho D, Jin J (2004) Higher criticism for detecting sparse heterogeneous mixtures. Ann Stat 32: 962-994.

14. Zaykin DV, Zhivotovsky LA, Czika W, Shao S, Wolfinger RD (2007) Combining p-values in large scale genomics experiments. Pharm Stat 6: 217-226.

15. Yu K, Li Q, Bergen AW, Pfeiffer RM, Rosenberg PS, et al. (2009) Pathway analysis by adaptive combination of P-values. Genet Epidemiol 33: 700-709.

16. Hess A, Iyer H (2007) Fisher's combined p-value for detecting differentially expressed genes using Affymetrix expression arrays. BMC Genomics 8: 96.

17. Li S, Williams BL, Cui Y (2011) A combined p-value approach to infer pathway regulations in eQTL mapping. Stat Interface 4: 389-402.

18. Zaykin DV, Kozbur DO (2010) P-value based analysis for shared controls design in genome-wide association studies. Genet Epidemiol 34: 725-738.

19. Dai H, Charnigo R (2007) Omnibus testing and gene filtration in microarray data analysis. J Appl Stat 34: 1165-1183.

20. Dudbridge F, Koeleman BPC (2004) Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. Am J Hum Genet 75: 424-435.

21. Greene CS, Penrod NM, Kiralis J, Moore JH (2009) Spatially uniform relieff (SURF) for computationally-efficient filtering of gene-gene interactions. BioData Min 2: 5.

22. Moore JH, White BC (2007) Tuning ReliefF for Genome-Wide Genetic Analysis. Lecture Notes in Computer Science 4447: 166-175.

23. Robnik-Sikonja M, Kononenko I (2003) Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learning Journal 53: 23-69.